Audio Engineering Society
Convention
Paper

Presented at the 116th Convention
2004 May 8–11 Berlin, Germany

Spatial decomposition of time-frequency regions: subbands or sinusoids

Aki Härmä[1] and Christof Faller[2]

[1]*Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, P.O. Box 3000, 02015, HUT, Finland*

[2]*Mobile Terminals Division, Agere Systems, Allentown, PA, USA*

Correspondence should be addressed to Aki Härmä (Aki.Harma@hut.fi)

ABSTRACT

Techniques where a stereo or a multichannel signal is decomposed into spatial source-labeled time-frequency slots by level, time-difference, and coherence metrics have become popular in recent years. Good examples are binaural cue coding and up/downmixing techniques. In the article, we will provide an overview and discuss parallel approaches in the field of array processing and blind source separation. Typically, time-frequency slots are formed from subband representations of signals. However, it is also possible to produce a similar spatial decomposition for a parametric representation (sinusoids, transients, and noise) of a stereo or multichannel audio signal. Advantages and disadvantages of the two approaches for audio coding applications are discussed in this article.

1. INTRODUCTION

In some cases a stereo or a multichannel audio signal is composed of independent channels such as separate tracks of a studio recording. In these recordings it is not generally meaningful to characterize the set of signals by common spatial attributes such as directions of sources or a spatial image. Another class of multichannel recordings is produced from a recording with an array of microphones, or a mix produced by amplitude panning or other rendering techniques. In this article we mainly consider the latter class of stereo or multichannel signals.

In these multichannel signals individual sources coexist in different channels of the recording. In a general case a P-channel audio signal $\mathbf{X}$ is produced from N independent source signals $\mathbf{S}$ by the following matrix expression

$$\mathbf{X} = \mathbf{M}\mathbf{S}, \tag{1}$$

where $\mathbf{M}$ is a $P \times N$ matrix. For example, a stereo signal $X = [x_l(t)\ x_r(t)]^T$ may be produced from an array of original source signals $S = [s_0(t)\ s_1(t)\ \cdots\ s_{N-1}(t)]^T$ by a $N \times 2$ scalar amplitude panning matrix $\mathbf{M}$, where each column has a pair of gain factors for each source signal. When the multichannel signal is captured using an array of microphones in a recording space, the matrix formulation of (1) applies when the signals in matrices $\mathbf{S}$ and $\mathbf{X}$ are replaced by their respective Fourier transforms and each element of $\mathbf{M}$ is the Fourier transform of the acoustic transfer function from a source location to a microphone. In more complex scenarios $\mathbf{M}$ can also be time-varying.

In both examples it is meaningful to say that $\mathbf{M}$ contains all information about the spatial representation of the observed multichannel signal $\mathbf{X}$. In *spatial decomposition* of a multichannel signal, the aim is to find $\mathbf{M}$ given an observed set of signals $\mathbf{X}$. In *source separation*, the goal is to find original sources signals $\mathbf{S}$ from signals $\mathbf{X}$. In this article the problem of source separation is considered to be a part of the problem of spatial decomposition. Spatial decomposition is useful in many different applications. In audio coding, spatial decomposition of a multichannel signal can be as useful as frequency decomposition, that is, it can be used to allocate more bits to spatial regions where they are more needed. A spatial decomposition can also be used to manipulate or remix a recording by changing directions and levels of individual sources. For example, in a teleconference application individual speakers could be amplified or attenuated as desired. A spatial decomposition can also be used to adapt to a new loudspeaker configuration [1, 2], or in different types of enhancement, suppression, and re-panning applications [3].

Finding unknown source signals $\mathbf{S}$ and the mixing matrix $\mathbf{M}$ from signals in $\mathbf{X}$ is an inverse problem which is impossible to solve without regularization of the problem. In blind source separation (BSS) it is required that the source signals are independent [4] and the number of source signals is the same or less than the number of channels in $\mathbf{X}$. It may also be assumed that signals $\mathbf{X}$ are recorded with a specific microphone array [5]. In fact, even if $\mathbf{M}$ is known source separation is not perfect unless the source signals are independent and $\mathbf{M}$ can be inverted. Clearly, the general formulation of the problem is impossible to solve exactly, other than in uninteresting trivial cases (e.g., when $\mathbf{M}$ is an orthogonal transform matrix).

However, in many applications a mathematically exact spatial decomposition is not necessary. In audio coding very simple techniques such as sum-difference coding [6] does already a good job in reducing the bitrate of a stereo signal. Sum-difference coding is basically an application of beamforming techniques to the problem of spatial decomposition of a stereo signal. Here the sum signal corresponds to the sources panned to the middle (or the median plane) and the difference has sources spatially at the sides in the original stereo signal [also called mid-side (M/S) coding]. This generalizes to Walsh-Hadamard transform coding or actually the use of any orthogonal transform matrix applied to samples or subband samples of a multichannel signal [5]. Here the spatial decomposition of a multichannel signal is performed with a fixed transform matrix $\mathbf{M^{-1}}$. The transform matrix can also be signal dependent such as in Karhunen-Loeve Transform (KLT), which has been applied to multichannel audio coding [7] and upmixing applications [1] for adaptive spatial decomposition of fullband signals.

A simple way to estimate some properties of the mixing matrix $\mathbf{M}$ adaptively from signals is to continuously measure level-differences and time-differences between the signals. This can be used as side information in coding or as information for re-panning of signals. Intensity stereo coding is based on this principle but it is applied separately at different subbands of a subband audio coder [8]. The binaural cue coding (BCC) method introduced in [9] is doing basically the same but the time-frequency decomposition of stereo or multichannel signals is based on FFT and is separate from the MDCT

used in the coding algorithm. A similar audio coding algorithm has been recently introduced in the context of parametric coding [10]. In their approach level and time differences, and coherence between the channels were estimated from a subband decomposition very similar to BCC. In this article we study a method where the spatial decomposition is estimated separately for sinusoidal components.

The goal is to estimate the mixing properties separately for different components of a decomposed signal rather than for samples or frames of fullband signals. The wisdom in this approach is that in non-stationary audio material with multiple sources there are usually *simple* time-frequency regions with only one dominating source. In such regions we can make a successful spatial decomposition. In time-frequency regions where we fail, we just try to do something which sounds tolerable for a particular application. We may safely argue that in many applications the time-frequency regions where a spatial decomposition method fails due to complexity are the regions where our spatial hearing mechanism will also fail in acquiring reliable spatial cues.

There are infinitely many different (time-frequency) decompositions of multichannel signals that are in principle applicable. In this article we compare two different approaches. Firstly, we review the binaural cue coding method introduced in [9] which is based on a subband decomposition of signals. Secondly, we introduce a method where essentially the same *cues*, that is level and time differences, are estimated from sinusoidal decompositions of signals.

In the current article we limit the discussion to a special case of stereophonic audio recordings. In addition, the presented techniques are based on specific regularization of the estimation problem, i.e., it is assumed that stereo signals are generated by amplitude panning of original source signals. That is, we assume that each component of a signal has a scalar mixing matrix $\mathbf{M}$ that we are trying to estimate.

2. SPATIAL CODING IN SUBBANDS

The frequency decomposition used in BCC (and other related approaches) is motivated by the fact that the auditory system has a limited spectral resolution. It is assumed that, given a mono sum signal, any spatial image containing the sources in the sum signal can be rendered by synthesizing appropriate spatial cues in a number of subbands. A BCC encoder and decoder based on this assumption are shown in Fig. 1. The BCC encoder estimates the inter-channel cues which are the determining factors for the perceived spatial image of the input multi-channel audio signal. These cues are quantized, coded, and transmitted to the BCC decoder along with a single downmixed audio channel. Given the transmitted single channel and the transmitted inter-channel cues, the BCC decoder generates a multi-channel signal with cues approximating the cues of the original multi-channel signal.

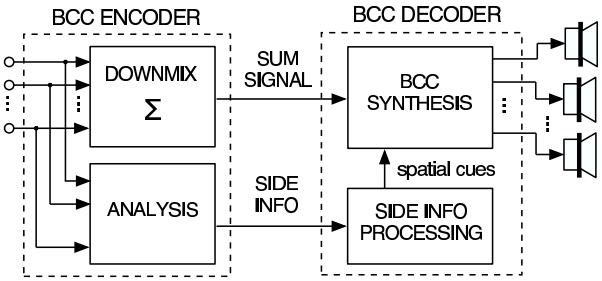

Fig. 1: Generic encoder and decoder scheme of BCC

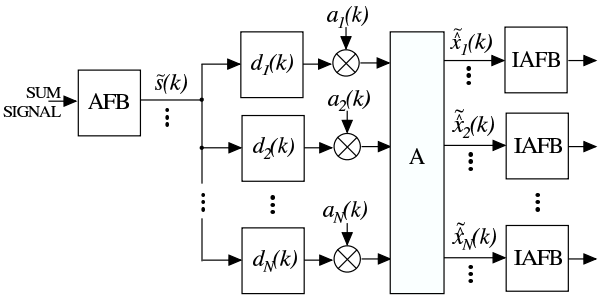

Fig. 2: Detailed scheme for multi-channel synthesis given the transmitted single audio channel.

Figure 2 shows the details of the "BCC synthesis" block of Fig. 1. The transmitted audio channel is decomposed into subbands by an auditory filterbank (AFB). AFB here denotes an invertible filterbank with subbands with a bandwidth equal or proportional to the critical bandwidth of the auditory system. Time differences between channel pairs are synthesized by imposing delays on the subband signals and level differences are synthesized by applying different gain factors. The processing block A in Fig. 1 is a mechanism to reduce the coherence between the subband signals of the output channels. This is used for synthesizing coherence cues [10, 11].

In this study we are focusing on the properties of level

difference cues only.

3. SINUSOIDAL SPATIAL CODING

Sinusoidal or parametric representations of speech [12, 13] and audio signals [14] have been developed and studied by many authors. Parametric representation of audio signals is known to be a very useful domain for manipulation of audio material, see e.g. [15] for a review.

The sinusoidal coder used in this article is based on a parametric line spectrum estimation method which is often called Analysis-By-Synthesis/Overlap-Add (ABS/OLA) when referring to an efficient frequency-domain algorithm proposed by George and Smith [16]. The only difference to the original algorithm is that here the signal envelope normalization is omitted.

The algorithm subtracts iteratively for $k = 0, 1, \cdots, K$ windowed sinusoidal pulses,

$$s_k(t) = A_k \cos(\omega_k t + \phi_k) w(t)\,, \tag{2}$$

from the residual $e_k(t)$ such that

$$|e_k(t) - s_k(t)|^2 \tag{3}$$

is minimized at each step. For $k = 0$ the residual corresponds to the windowed original signal. The window function $w(t)$ is applied to the original signal with 50% overlap. The estimation technique is such that the frequency of a highest spectrum peak is chosen from the spectrum of $e_k(t)$ at each iteration. Amplitude A_k and phase ϕ_k terms are then computed such that they minimize (3) (see [16] for more details). Next, e_{k+1} is computed by

$$e_{k+1}(t) = e_k(t) - s_k(t) \tag{4}$$

to produce a signal frame where the sinusoidal pulse has been removed. The last step can be implemented very efficiently in the FFT domain. The estimated sinusoidal signal can be synthesized from sinusoidal pulses directly with overlap-add.

3.1. Decomposition of a multichannel signal

For a stereo or multichannel signal we consider a simple modification of the algorithm. P signals in a matrix $E_k = [e_{k0}\ e_{k1}\ \cdots\ e_{k(P-1)}]^T$ are synchronously windowed with the Hann window. For each frame of a multichannel signal and at each iteration the one highest peak of all spectra is identified and its frequency ω_k is determined. Then amplitude and phase terms, A_{kp} and ϕ_{kp}, at frequency ω_k are computed for all signals $p = 0, 1, \cdots, P-1$. The residual for the next iteration is then given by

$$e_{(k+1)p}(t) = e_{kp}(t) - A_{kp} \cos(\omega_k t + \phi_{kp}) w(t), \tag{5}$$

for $p = 0, 1, \cdots, P-1$. Since the algorithm is based on iterative subtraction of K estimated sinusoids from the P channels it holds that perfect reconstruction of an original windowed signal frame at channel p can be obtained with

$$e_{0p}(t) = e_{Kp}(t) + \sum_{k=0}^{K-1} A_{kp} \cos(\omega_k t + \phi_{kp}) w(t)\,. \tag{6}$$

The result of the decomposition is a matrix of residual signals E_{K-1}, an array of frequency terms ω_k ($k = 0, \cdots, K-1$) common to all signals, and amplitude and phase terms A_{kp} and ϕ_{kp}, respectively, for each channel and sinusoid. Note that the frequency of a sinusoidal pulse is the same in all audio channels. Therefore, this scheme is a model of a set of *spatial sinusoids*, which a much more compact representation than a scheme where a sinusoidal model is applied separately to different audio channels.

3.2. Spatial decomposition by sinusoids

The spatial decomposition of the multichannel signal can now be performed by the analysis of amplitude and phase terms A_{kp} and ϕ_{kp}, respectively. For example in amplitude panned stereo signals, level differences of a k^{th} sinusoidal pulse can be expressed by

$$L_k = 20 \log_{10} \left(\frac{A_{kl}}{A_{kr}} \right), \tag{7}$$

where A_{kl} and A_{kr} are amplitudes of sinusoids at the frequency of ω_k in the left and right channels, respectively.

In order to separate sinusoids corresponding to an original source signal which has been panned to the left side of a spatial image we may pick only sinusoids for which $L_k > 0$ and attenuate others. In purely amplitude panned stereo signal this will give high attenuation for signals panned to the right. Similarly, time-difference between sinusoidal pulses at ω_k may be estimated by computing the phase delay given by

$$d_k = (\phi_{kl} - \phi_{kr})\omega_k. \tag{8}$$

However, this is difficult especially at high frequencies because of the ambiguity of the phase term.

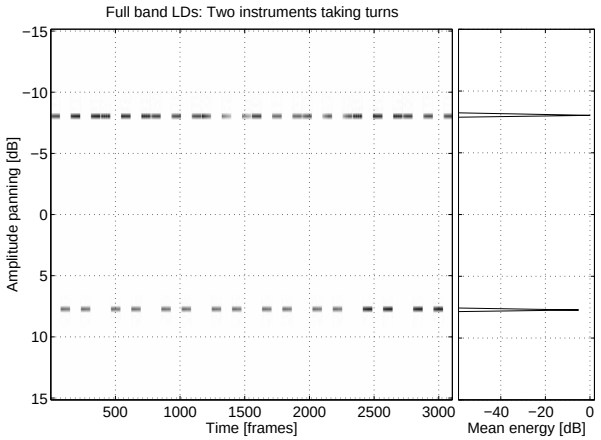

Fig. 3: A pangram for an amplitude panned stereo signals where two instruments are taking turns in playing a melody. This pangram was computed from full-band signal energy in frames of 1024 samples. The two instruments were panned to 8dB left and right, respectively.

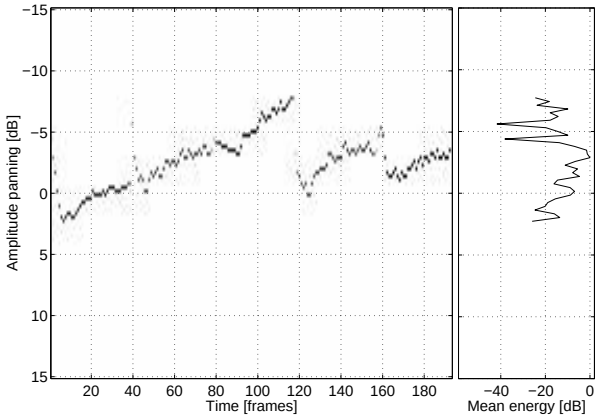

Fig. 4: A pangram for an amplitude panned stereo signal where two instruments (cello and clarinet) are playing synchronously a different melody. This pangram was computed from full-band signal energy in frames of 2048 samples. The two instruments were panned to 8dB left and right, respectively.

The residual signal is still a multichannel signal. It has a low energy if the sinusoidal model is successful in modeling the original signals but it is typically non-vanishing and absolutely necessary in high quality audio applications. The coding of the residual signals is discussed later in this article.

4. EXPERIMENTS

Estimates of amplitude panning information in a stereo signal can be visualized using a graph showing the distribution of signal energy being panned to different directions. In this article, such a graph is called a *pangram*. Figure 3 gives a simple example. The x-axis in the left panel is time in frames and the y-axis represents the amount of signal energy panned to different directions. In this example the original signal was a music signal where two instruments take turns such that there is no temporal overlap. The two instruments were amplitude panned with 8dB and -8dB to the left and right in the stereo signal. Since the two source signals in Fig. 3 do not overlap in time, level differences can be estimated directly by measuring RMS values from time frames of fullband signals. In this case, estimation of level differences from a fullband signal, time-frequency slots of a subband coder, or sinusoidal decomposition of the signals give almost identical results.

In order to compare the difference between different decompositions of signals we created a collection of music signals where there are two amplitude panned (at -8 and 8 dB) instruments playing synchronously a different melody. Some of these samples are available on our Internet site [17]. In the following, these signals are called *Set I*. The set has 21 sequences with all combinations between cello, clarinet, french horn, saxophone, flute, violin, and piano. The set was assumed to be very difficult because the notes have been carefully aligned in time so that they start and end at the same time and they have a similarly rich spectral structure. For these signals the estimation of the pangram from signal energies of a fullband signal almost completely fails, as is illustrated for the cello-clarinet pair in Fig. 4.

The pangram produced using the BCC algorithm is shown in Fig. 5. It gives peaks around -8 dB and 8 dB panning directions where the original sources were panned to. But the peaks are often shifted towards the center and there are many false peaks. The shifting of peak positions and false peaks may cause an effect which is sometimes encountered in BCC synthesis: sources are not spatially placed in static locations but they fluctuate in time [18]. If the panning information depicted in Fig. 5 is used to control the resynthesis of a stereo signal from a downmixed monophonic signal, fluctuations of the left (upper) source may be expected.

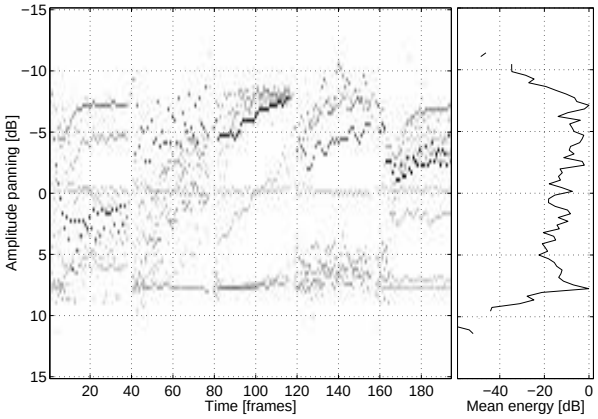

Fig. 5: A pangram for the same signal as in Fig. 4 but computed using the BCC analysis, that is a subband decomposition of the stereo signal. The length of the analysis FFT window is 2048 samples and each complex spectrum has been divided into 20 nonuniform approximately 2 ERB wide frequency bands for the analysis.

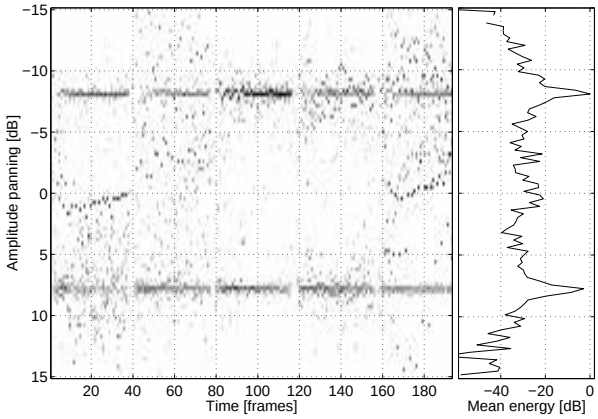

Fig. 6: A pangram for the same signal as in Fig. 4 but computed for a sinusoidal decomposition of the stereo signal. The length of the analysis window is 2048 samples and each frame has been modeled using 20 sinusoids.

Figure 6 shows the pangram produced from the amplitude parameters A_{kl} and A_{kr} of a sinusoidal model estimated from the same cello-clarinet duet as in Figs. 4 and 5. The total number of sinusoids in each frame of 2048 samples was 20. The cues of the two sources are clearly implied by the two peaks, although there is also a significant amount of errors (values between peaks). Those errors are produced by overlapping harmonics in the two instrument signals. The right panel in Figs. 6 and 4 is a plot of the mean energy distribution computed from the pangram. In sinusoidal decomposition the source directions at -8 dB and 8 dB give peaks which are 20 dB higher than the middle region around the 0 dB panning direction. In the BCC case, however, the difference between correct peak positions and the region in between is much smaller.

The difference between the two methods results from the fact that the frequency resolution is lower in the subband scheme of BCC than in the sinusoidal model. In Figs. 6-5, the length of the analysis window was 2048 samples. In fact, if the length of the window was 512 or 1024, the difference between the sinusoidal and BCC cases would be significantly reduced.

Another set (*Set II*) of ten test signals was produced by panning randomly four monophonic source signals to four directions (amplitude differences of -8, -4, 4, and 8 dB). The source signals contained many different samples of music and speech sequences including percussive sounds (e.g., castanets). This data set represents typical stereo audio material where spectrum overlap is not as severe as in Set I. The mean energy distribution estimated using the BCC method and sinusoidal modeling with 20 sinusoids are shown in Fig. 7a. The dotted curve in the middle represents the spatial energy distribution of the residual signal of the sinusoidal model computed using the BCC algorithm. We may now define two measures (see Fig. 7a) which can be used to characterize the performance of the two algorithms. The *modeling gain*, $G_s[\mathrm{dB}]$ is similar to the classical prediction gain and simply gives the difference between the original signal and the residual signal after subtraction of sinusoids. This is averaged over all panning directions. The difference between a peak value in a mean energy distribution curve and a local minimum between maxima is here called *panning gain*, $G_p[\mathrm{dB}]$. This is illustrated in the bottom curve of Fig. 7a.

Figure 7b shows the modeling gain and panning gain defined in Fig. 7a averaged over the set of four-source signals. Increasing the number of sinusoids obviously increases the modeling gain, that is, the energy of the residual decreases. However, the difference between the sinusoidal coder and BCC algorithm in terms of the panning gain G_p decreases as the number of sinusoids grows and

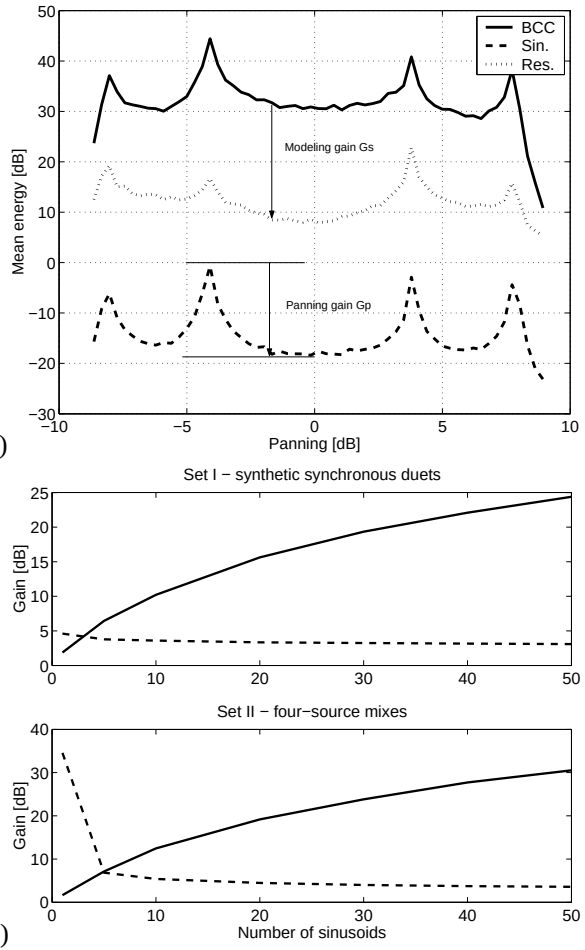

Fig. 7: a) The mean energy distribution averaged over the Set II of stereo signals with four sources. The curve for the sinusoidal case (dashed) was shifted down for illustative reasons. b) Modeling gain G_s in sinusoidal modeling (solid) and the average difference in panning gain G_p (dashed) between sinusoidal modeling and BCC in Set I (top) and Set II (bottom) signals. In both algorithms, the length of the analysis window was 1024 samples.

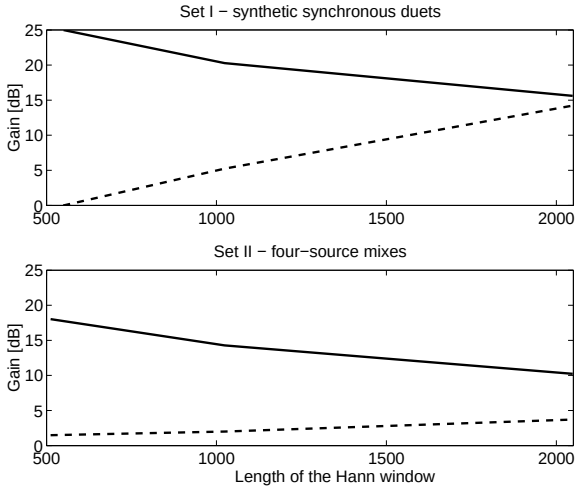

Fig. 8: Modeling gain G_s in sinusoidal modeling (solid) and the average difference in panning gain G_p (dashed) between sinusoidal modeling and BCC in Sets I and II signals as a function of the length of the analysis window.

finally, in this particular case, remains approximately at the level of 4 dB. This indicates that after a certain point the sinusoidal modeling starts losing its efficiency and extracted sinusoidal pulses represent noise rather than spectrum peaks. The same trend can also be seen using signals from Set I (top panel). There, both gain values are approximately 3-5 dB higher than in bottom panel of Fig. 7b but the overall trend is the same.

The length of the analysis frame may change the results significantly. The modeling gain and the difference in panning gain in the two algorithms is illustrated in Fig. 8. In all window sizes the number of sinusoidal components was 20. When the analysis frame is short, the modeling gain will be higher because the sinusoidal model can model finer temporal details in the signals. But, increasing the length of the analysis window improves the panning gain in sinusoidal modeling. For example, in the Set I simulations in the top panel of Fig. 8 the difference in panning gain between the sinusoidal and BCC cases is almost 15 dB for a window length of 2048 but nearly zero for a 512-sample Hann window.

5. DISCUSSION

In the set of signals in Fig. 7, the spatial energy distribution of the residual signal (dotted curve) still has peaks in the same positions as in the original signal, although

the difference between a peak and a valley is somewhat smaller. That is, the spatial distribution of original sources largely remains in the residual signal even if the largest sinusoidal components have been removed from the signals (this is also easy to hear [17]). Therefore, it is meaningful to consider also making a spatial decomposition for the residual signal.

Let us consider a hybrid system where the modeling of amplitude panned sinusoids is followed by BCC-type subband processing of the residual stereo signal. The first phase would be the sinusoidal parametrization of the stereo signal. It would be beneficial to perform sinusoidal modeling at different time resolutions adaptively according to the momentary signal properties. In addition the number of sinusoids could also be chosen adaptively in each signal frame, e.g., using a similar stopping criterion as has been proposed in [19]. After the subtraction of sinusoids the residual is processed using subband coding similar to the BCC algorithm. The residual signal can be coded with a significantly lower bitrate than the original audio signal. Using the classical coding theoretic approximation given by $\mathrm{SNR} \approx 6\mathrm{bits} + \gamma$ we may argue that the bitrate can be reduced by 2 bits/sample if the modeling gain is 12 dB. This margin would be sufficient for the coding of the sinusoidal data because typically the sinusoidal components in one audio channel can be coded with less than 20 kbits/s [20]. Therefore, we may anticipate the proposed coder could produce equal or better quality of a stereo signal at a similar or a slightly smaller bitrate than a subband coder based on BCC only.

In re-panning and up/downmixing applications the better spatial resolution of the sinusoidal approach combined with subband processing of the residual may also improve the performance compared to pure subband processing.

In the current article, we only studied spatial decomposition of a stereo signal based on level differences between the two channels. It is clear that in both algorithms we can also estimate the time differences between sinusoids or subbands of the two channels.

6. CONCLUSIONS

In this article we have studied two different approaches to acquire spatial information from a stereo audio signal by comparing amplitudes of different time-frequency components of the signals. Spatial decomposition of a stereo or multichannel signal is beneficial in many different applications such as audio coding, up/downmixing, and equalization or manipulation of a stereo image.

The first algorithm is based on subband decomposition where amplitude differences between the two stereo signals has been estimated individually in each subband. In the second algorithm, the stereo signal is divided into a set of sinusoidal components where we may estimate a spatial decomposition of the stereo signal from amplitudes of sinusoidal components in the two channels.

The results presented in this article were based on the use of a representation which shows how signal energy is distributed in terms of amplitude panning to different directions in the spatial image. In general, we may argue that the spatial information related to amplitude panned sources can be estimated more accurately from a sinusoidal representation of a stereo signal than from the subband representation. In some cases the difference is small. In particular, when the analysis frame is short (512 sample), the two algorithms produce almost similar results. However, the benefits of the sinusoidal approach become very clear if the analysis window is allowed to be long (2048 samples).

7. ACKNOWLEDGEMENTS

The work of A. Härmä has been supported by the Academy of Finland.

8. REFERENCES

[1] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *J. Audio Eng. Soc.*, pp. 914–926, November 2002.

[2] C. Avendano and J. M. Jot, "Frequency-domain techniques for stereo to multichannel upmix," in *Proc. 22nd AES Int. Conf. on Virtual, Synthetic, and Entertainment Audio*, (Espoo, Finland), pp. 121–130, June 2002.

[3] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY, USA), pp. 55–58, October 2003.

[4] D. Yellin and E. Weinstein, "Criteria for multichannel source separation," *IEEE Trans. Signal Processing*, vol. 42, pp. 2158–2168, August 1994.

[5] A. Härmä, "Coding principles for virtual acoustic openings," in *Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, (Espoo, Finland), June 2002.

[6] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, pp. 569–572, 1992.

[7] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "An intra-channel redundancy removal approach for high-quality multichannel audio compression," in *109th AES Convention, Preprint 5238*, (Los Angeles, USA), p. 14, September 2000.

[8] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *96th AES Convention, Preprint 3799*, February 1994.

[9] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE Workshop Appl. Signal Processing, Audio and Acoust.*, (New Paltz, New York, USA), September 2001.

[10] E. Schuijers, W. Oomen, and B. den Brinker, "Advances in parametric coding for high-quality audio," in *AES 114th Convention preprint*, (Amsterdam, The Netherlands), March 2003.

[11] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and applications," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, November 2003.

[12] P. Hedelin, "A tone-oriented voice-excited vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. I, (Atlanta, USA), pp. 205–208, IEEE, March 1981.

[13] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal speech model," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 34, pp. 744–754, August 1986.

[14] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for nonharmonic sounds based on sinusoidal representation," in *Proc. Int. Computer Music Conf.*, pp. 290–297, 1987.

[15] M. Goodwin, M. Wolters, and R. Sridharan, "Post-processing and computation in parametric and transform audio coders," in *Proc. 22nd AES Int. Conf. on Virtual, Synthetic, and Entertainment Audio*, (Espoo, Finland), pp. 149–158, June 2002.

[16] B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 389–406, September 1997.

[17] A. Härmä and C. Faller, "Audio demostrations for the 116th aes convention paper." http://www.acoustics.hut.fi/˜aqi/aes116demo, May 2004.

[18] F. Baumgarte and C. Faller, "Why Binaural Cuse Coding is better than Intensity Stereo Coding," in *112th AES Convention Preprint*, (Munich,Germany), May 2002.

[19] N. H. van Schijndel, M. Gomez, and R. Heusdens, "Towards a better balance in sinusoidal plus stochastic representation," in *Proc. Workshop Appl. Signal Processing to Audio and Acoustics (WASPAA, 2003)*, (New Paltz, NY, USA), pp. 197–200, October 2003.

[20] A. C. den Brinker, E. G. P. Schuijers, and A. W. J. Oomen, "Parametric coding of high-quality audio," in *112th AES Convention preprint 5554*, (Munich, Germany), May 2002.